MACHINE LEARNING CLUSTERING IN FINANCIAL MARKETS: A LITERATURE REVIEW

Ștefan RUSU, Marcel BOLOȘ

University of Oradea, Oradea, Romania
stefan.rusu.d@gmail.com
marcel_bolos@yahoo.com

Abstract: *This paper aims to present a concentrated overview of innovative research in machine learning clustering techniques as applied to to different facets of financial markets and stock market investing. Research on techniques such as K-Means Clustering or Agglomerative Hierarchical Clustering and their derivatives play a pivotal role in augmenting the stock market research and investment strategies. Wheter it is time series clustering and prediction, portfolio selection and optimization, or risk management, machine learning clustering has potential to enhance already existing processes by improving performance, reducing time spent on repetitive tasks or mitigating human errors. A truly innovative tool in the investor's toolset, it is imperative to not overlook its limitations, such as the necessity of selecting the appropriate technique for specific datasets, or the need for human supervision to maximize its utility and insights extracted.*

Keywords: *artificial intelligence; machine learning; clustering; finance; investing; financial markets*

JEL Classification: C38; C53; C55; C61; C8; G10; G11; G17

1. Introduction

With higher availability of data and increase in the computing power experimented in the last years, Artificial Intelligence (AI) and Machine Learning (ML) have started to become re-enter the sphere of interests for many domains, scientific or commercial alike. It is safe to assume that the rise in popularity of Open AI's Generative Pre-trained Transformer, ChatGPT, has provided the spark that reignited the general public's interest to AI as well.
Within the ML subset of AI, the clustering subset of techniques and its use in financial markets is the focus of this paper. Clustering refers to the grouping of more observations in a way that similar observations end up in the same grouping called cluster. Clustering has many aplication in a plethora of domains, and the financial and investing fields are no exception.
Some of the use cases in which clustering can be used in financial markets are:

- *Portfolio construction*. Grouping similar stocks based on their risk and return profiles, or other relevant factors can prove useful in building diversified portfolios, thus reducing idiosyncratic risk.
- *Risk management*. Grouping stocks with similar risk profiles help investors a better understanding of similar-risk stocks.
- *Asset allocation*. Having clusters based on asset classes or sector provenience, investors can avoid including similar equities in their portfolios and reduce their overall concentration risk.
- *Market analysis*. Clustering different macro regimes and business cycle phases can prove useful in adapting the investing style to different circumstances to maximize returns and/or to minimize losses.

- *Factor identifiation*. Having clusters based on several attributes of stock market companies, an investor can identify the factors that drive the stock's movements and capitalize on them.
- *Trading strategies identification*. By clustering equities that display similar patterns, investors can develop trading strategies to obtain better returns or follow specific objectives.
- *Investor segmentation*. Clustering different types of investors (such as large institutional investors and individual investors) and their behavior can aid the generation of useful information about the positioning of other investors in the stock market.
- *Anomaly detection*. Having equities clustered by specific measures, investors can more easily see the equities that display abnormal behaviors and react accordingly.
- *Other investment-related research*. Considering that the number of factors that influence the movement of publicly traded assets is so enormous that is near-impossible to quantify, using clustering techniques on different metrics that the investor thinks might have an impact on the asset's prices can be a great tool to validate or invalidate these hypotheses.

In this paper, a literature review of research done in the field of ML clustering, with applications to the financial market is presented to highlight various techniques that can be used to improve or simplify various aspects of the stock investment process.

2. Literature review

Madhulatha (Madhulatha, 2012) provides a description of clustering as a widespread technique employed in diverse domains, such as machine learning, data mining, pattern recognition, image analysis, and bioinformatics. The author states that clustering involves the act of categorizing similar objects or dividing a dataset based on a specified distance metric. The heavy dependance on the analyst's decision is emphasized here. It is suggested that combining diverse clustering results yield more robust clusters that are not excessively reliant on criteria. Specific constraints that are linked to clustering are acknowledged as well. The study advises against generalizing clustering outcomes beyond the examined dimensions or variables, because clustering consistently generates clusters even in the absence of an intrinsic group configuration in the data.
Rani and Sikka (Rani & Sikka, 2012) offer a comprehensive survey and synthesis of previous research on time-series clustering. The use of time-series clustering plays a crucial role in the field of data mining, as it facilitates understanding of the underlying mechanisms responsible for the generation of time-series data and enables prediction of future values. The paper underscores the varied selection of application domains that have been investigated in relation to time-series clustering, encompassing fields such as science, engineering, business, finance, economics, healthcare, and government. The conclusion highlights the growing interest focused on the extraction of valuable insights from time series data, with a particular emphasis on the domain of time-series clustering. The careful assessment of proposed algorithms is essential due to the large size of time-series data and the possible existence of outliers.
Cai et al. (Cai, et al., 2016) highlight the increasing significance of financial data analysis and the use of clustering techniques within this field. The authors underscore the lack of research that is specifically centered on clustering techniques for the purpose of analyzing financial data. By assessing various clustering techniques in the context of analyzing diverse financial data sets, encompassing both time series and transactions, the objective of the study is to improve understanding related to the internal framework of financial datasets

while assessing the proficiency of individual clustering techniques within this field. The article assesses various clustering algorithms and evaluates their effectiveness by utilizing metrics such as the Dunn Index (DI) and Davies-Bouldin Index (DBI). The paper underscores the importance of using normalized centroid-based clustering in the context of financial data analysis, while also drawing attention to the constraints of current methodologies. The study suggests that density-based clustering is an unsuitable approach for financial datasets. Normalized centroid-based clustering with a higher Dunn Index (DI) or a lower Davies-Bouldin Index (DBI) has been determined to be the optimal approach for determining the appropriate number of clusters for financial data classification. The authors recognize the limitations associated with clustering techniques, including the tendency of K-means to identify spherical clusters and the limited effectiveness of centroid-based clustering in handling noise. Thus, exploration of alternative centroid-based clustering techniques that are tailored for financial data sets is proposed.

Puerto et al. (Puerto, et al., 2020) present an innovative methodology that integrates clustering techniques and portfolio optimization to efficiently tackle the issue of portfolio selection in the field of finance. The proposed framework integrates a clustering-based criterion into the process of portfolio selection, providing a comprehensive solution that clusters assets and selects the optimal portfolio simultaneously. The key novelty of the approach resides in its use of a metric founded on correlation coefficients among the returns of assets to guide the clustering and selection process. The use of mixed-integer linear programming enables the model to efficiently distribute investment capital, considering both the cluster structure and the trade-off between risk and return. The integrated clustering and portfolio selection framework proposed provides various benefits in contrast with traditional two-step methodologies. The combination of clustering and selection into a single phase optimizes the decision-making process and improves effectiveness. Moreover, the framework's flexibility enables the integration of diverse risk measures and correlation metrics, rendering it adjustable to a range of investment tactics and inclinations.

Carta et al. (Carta, et al., 2021) introduced a methodology for real-time event detection by combining traditional newswires and microblogging platforms. This approach enables the extraction of significant information related to events across diverse domains. The algorithm employs clustering techniques to identify interrelated news articles from credible sources and integrates them with microblogging content to assess the significance of events in the public's perception. This study is centered on the financial sector and aims to verify the effectiveness of the methodology using empirical data obtained from Dow Jones, Stocktwits, and the S&P 500 index. The assessment showcases the proficiency of the approach in precisely identifying significant incidents, encompassing notable instances such as the Brexit Referendum and the Covid-19 pandemic. The methodology involves utilizing domain-specific representations of textual data, employing a hierarchical clustering technique, and eliminating outliers to produce significant groups of events. The proposed method is suitable for the detection of events in real-time situations and can be conveniently customized for diverse application domains with minimal modifications.

Vijayalakshmi Pai and Michel (Vijayalakshmi Pai & Michel, 2009) focus on the issues of diversification in small portfolios, as well as portfolio optimization problems that encompass a range of constraints, such as fundamental, bounding, cardinality, and class constraints. The authors highlight the need for an evolution strategy-based approach as a potential solution for addressing portfolio optimization challenges that involve various constraints. The employed methodology utilizes k-means cluster analysis as a means of simplifying the model and eliminating the cardinality constraint. The research shows that portfolio sets clustered through k-means offer dependable outcomes by displaying proximity to the exact efficient frontier. Additionally, this approach encourages dimensionality reduction, which leads to faster convergence. In terms of efficiency and robustness, the proposed evolution strategy shows superior performance when compared to other methods such as generational genetic algorithms and quadratic programming. The proposed strategy is well-suited for small portfolios seeking diversification, especially in cases where there are constraints within asset classes.

Ahmed et al. (Ahmed, et al., 2020) provide a thorough overview of the k-means clustering algorithm, emphasizing its constraints and the need for improvement in the field of financial asset analysis. This study examines crucial concerns related to the k-means algorithm, such as the random initialization of centroids, the need for predetermined cluster numbers, and the algorithm's inability to manage diverse data types that are frequently encountered in financial asset analysis, with the aim to improve the understanding of the k-means algorithm and identify potential areas for further research to overcome its limitations. This study distinguishes itself through an in-depth review of prior literature and the implementation of broad experimental evaluations on benchmark datasets, aimed at assessing multiple iterations of the k-means algorithm. The findings show that the challenges presented by the k-means algorithm do not have a universally applicable solution, and instead, various adaptations are customized to suit specific data types or applications. The authors highlight the importance of creating a robust k-means algorithm that can handle the obstacles posed by initialization and mixed data types in a unified manner.

Wu et al. (Wu, et al., 2012) introduced a novel framework for forecasting stock trends through the integration of sequential chart patterns, the K-Means algorithm, and the AprioriAll algorithm. The employed methodology involves the conversion of stock price sequences into visual representations using a sliding window technique, followed by the clustering of the resulting charts into distinct patterns via the application of K-Means clustering. The chart patterns are subject to a detailed analysis to extract recurring patterns using the AprioriAll algorithm. This approach facilitates the identification of market behaviors that are associated with trends and empowers the prediction of such trends. The method provides precise forecasts of trends and provides traders with signals for buying or selling to augment their predictions of stock prices. The system exhibits a favorable outcome in financial forecasting, resulting in additional gains and outperforming traditional investment strategies. Nevertheless, it is crucial to consider certain constraints of the research, such as the robustness of the suggested model under various market conditions, or the potential impact of external factors on its efficacy.

Bini and Mathew (Bini & Mathew, 2016) have proposed an analytical framework that employs data mining techniques to detect companies with higher profitability in the stock market. Their study centers on the use of clustering and regression as data mining methodologies. The authors conducted a comparative analysis of diverse clustering methodologies, namely partitioning-based, hierarchical, model-based, and density-based techniques, using validation indices. The findings suggest that the K-means algorithm, and the Expectation-Maximization algorithm, exhibit superior performance compared to the hierarchical and density-based techniques. Subsequently, the outcomes obtained from clustering are used in a multiple regression analysis to make predictions about future stock prices. The authors suggest the creation of an online stock prediction system as potential future research.

Bin (Bin, 2020) investigates a portfolio-building algorithm which uses a combination of the Sharpe ratio and K-means clustering on financial ratios to produce diversified stock market portfolios. Dissimilarity metrics, including historical stock price fluctuations, return on assets, and asset turnover ratios, are employed in the clustering of S&P 500 companies. The results suggest that portfolios generated through clustering techniques using stock price movements exhibit poor performance, whereas those constructed based on financial ratios exhibit consistent outperformance relative to the market average. The algorithm and Sharpe ratio are recognized to have certain drawbacks, such as their susceptibility to the investment horizon and assumptions about portfolio returns. The study's findings are subject to limitations due to the restricted dataset of only 232 companies, highlighting the need for further comprehensive data analysis.

Zhu and Liu (Zhu & Liu, 2021) present a novel financial risk indicator system that uses K-means clustering to identify early warning systems in the context of financial risk management. Besides its efficiency, simplicity and obiectivity, the K-means algorithm is advantageous due to its capacity to overcome subjective biases, estimate the likelihood and magnitude of risk occurrence, and facilitate risk management. The results indicate that the use of the K-means algorithm enhances the accuracy of risk forecasting, facilitates precise distinction among various risk levels, and establishes a scientific foundation for the mitigation and control of risks. However, the study lacks a thorough analysis of the effectiveness of the suggested financial risk early-warning system under various market conditions, as well as the ability of the approach to handle complex and large datasets.

Fang and Chiao (Fang & Chiao, 2021) investigate the use of the K-means clustering algorithm, along with an Artificial Fish Swarm Algorithm (AFSA) in the forecasting and recommendation of stocks. The K-means algorithm's sensitivity to initial parameters and local optimization issues can be tackled using the AFSA for optimization, which leads to the development of KAFSA. The authors further showcase the algorithm's precision in predicting stock performance. The reliability of the clustering results is supported by the analysis of financial metrics such as closing price, price-earning ratio, earnings per share, and return on net assets. The KAFSA algorithm has shown efficiency in the field of financial stock prediction and recommendation. It has been shown to provide accurate classifications and scientific recommendations to investors. However, the article's treatment of the scalability and robustness of the KAFSA algorithm in relation to larger datasets and its effectiveness in dynamic market conditions is not thoroughly examined. Furthermore, the authors do not explore the potential influence of outside factors on the accuracy of stock prediction, and it neglects to consider the computational complexity of the suggested methodology.

Creating an investment portfolio that optimizes returns while minimizing risk is a complex undertaking. Traditional portfolio theories frequently exhibit limitations in the automated identification of high-performing stocks from a vast array, while considering industry and market differences. In addition, there is a lack of consideration for risk aversion in the context of stock market downturns. Wu et al. (Wu, et al., 2022) introduced a novel approach to constructing portfolios, which capitalizes on the persistent trend features of the financial market. They use the K-means clustering algorithm for the purpose of clustering stocks according to their trend characteristics. The portfolios that have been created show superior returns and reduced risks in comparison to traditional portfolios. The approach involves screening stocks within individual clusters, in which those exhibiting the highest values of key metrics are chosen to form stock pools and portfolios. Moreover, the study determines the most favorable quantity of stocks in the portfolio via empirical analysis. The authors suggest the need for further research to improve the efficiency of clustering algorithms in the context of portfolio construction. This entails the investigation of alternative clustering techniques and the optimization of the approach to accommodate varied market conditions. This study contributes to the development of enhanced investment strategies by overcoming the constraints of traditional portfolio theories and using clustering algorithms.

Lee et al. (Lee, et al., 2010) introduce a novel clustering technique, namely HRK (Hierarchical agglomerative and Recursive K-means clustering), which aims to forecast near-term fluctuations in stock prices following the release of financial reports. The methodology entails the transformation of financial reports into feature vectors, followed by the application of hierarchical agglomerative clustering to divide them into distinct clusters. After clustering, recursive K-means algorithm is used to divide each cluster into smaller sub-clusters, in which representative feature vectors are selected for the purpose of prediction. The results indicate that the HRK approach is effective in forecasting near-term fluctuations in stock prices by utilizing financial statements. Furthermore, the incorporation of both qualitative and quantitative features in financial reports yields superior outcomes when compared to their individual analysis. The findings indicate the possibility of expanding the application of the approach to financial news articles, as well as integrating additional financial ratios, accounting elements, and technical indicators.

Yang et al. (Yang, et al., 2016) introduced a novel classification system, known as Business Text Industry Classification (BTIC), which aims to overcome the constraints of current industry classification systems. The authors underscore the importance of proficient industry classification in capital market research and acknowledge the constraints of current systems, including arduous manual processes and inconsistency. The subjective nature and slow adjustment of manual classification systems are emphasized by the authors while they introduce BTIC as an automated, innovative, and impartial alternative. The methodology includes the use of textual data extracted from corporate disclosures, with a specific focus on the business section of Form 10-Ks. The doc2vec document embedding technique is employed together with Ward's hierarchical clustering method to facilitate the categorization of the data. The results obtained by BTIC using financial ratios show its corresponding efficiency with commonly employed systems such as the Standard Industrial Classification (SIC) and the Global Industry Classification Standard (GICS) with regards to inter- and intra-sector homogeneity. The authors argue that BTIC outperforms SIC and GICS in various dimensions, such as process automation, objectivity, clustering flexibility, and result interpretability.
Nourahmadi and Sadeqi (Nourahmadi & Sadeqi, 2022) focus on traditional portfolio selection using cluster analysis. By using both k-means clustering and agglomerative clustering techniques to train four different models, the authors aim to create outperformant investment portfolios. The highest Sharpe ratio stocks from each cluster are chosen to create the portfolios. After developing four models and using various clustering techniques, the authors discovered that, according to silhouette scores and Sharpe ratios, agglomerative clustering with average linkage, trained on data from the previous year, generated the best results. The chosen portfolio outperformed random selection in terms of the Sharpe ratio, but it still lagged the market index. However, the portfolio outperformed the random selection and the average volatility of all clusters, with lower annualized volatility. The authors make several recommendations for improvement, including changing the training period, experimenting with various clustering parameters, and sharpe ratio-based portfolio optimization. Other performance evaluation metrics such as the Sortino ratio, Calmer, and Max Drawdown as well as qualitative analysis using industry data and time series analysis are also proposed as ways to improve the measurement of outperformance.

3. Conclussions

This paper presented several machine learning clustering techniques used for portfolio optimization, portfolio selection, time series clustering and prediction, identifying early warning systems in the context of financial risk management, analyzing diverse financial datasets, identifying higher profitability companies, or real-time event detection.
It can be seen that the use of ML clustering in the field of financial markets is of great interest to researchers and practitioners alike. Several aspects of the stock market investment process can be tackled either to provide improved performance, to save time spent doing repetitive processes, or to reduce human errors.
And although ML clustering is a powerful and versatile tool, it is not without its drawbacks. It is not a one-size-fits-all solution, and the most suitable techniques must be used with the datasets it would make the most sense and provide the highest improvement when used. And while it is a subset of artificial intelligence, this does not mean that these algorithms should be blindly ran unsupervised. There is still the need for human supervision and knowledge to analyze the results and extract the needed information from them.
These being said, ML clustering techniques can provide an efficient and versatile tool in every investor’s toolset, but some human knowledge and insights is still needed to use them to the full potential.

References

1. Ahmed, M., Seraj, M. & Islam, S., 2020. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics, 9(8), 1295.
2. Bini, B. & Mathew, T., 2016. Clustering and Regression Techniques for Stock Prediction. Procedia Technology, 24(), p. 1248–1255.
3. Bin, S., 2020. K-Means Stock Clustering Analysis Based on Historical Price Movements and Financial Ratios. s.l.:CMC Senior Theses.
4. Cai, F., Le-Khac, N.-A. & Kechadi, M.-T., 2016. Clustering Approaches for Financial Data Analysis: a Survey, s.l.: School of Computer Science & Informatics, University College Dublin, Ireland.
5. Carta, S. et al., 2021. Event detection in finance using hierarchical clustering algorithms on news and tweets. PeerJ Computer Science 7:e438.
6. Fang, Z. & Chiao, C., 2021. Research on prediction and recommendation of financial stocks based on K-means clustering algorithm optimization. Journal of Computational Methods in Sciences and Engineering, pp. 1-9.
7. Lee, A. J., Lin, M.-C., Kao, R.-T. & and Chen, K.-T., 2010. An Effective Clustering Approach to Stock Market Prediction. PACIS 2010 Proceedings.
8. Madhulatha, T. S., 2012. An Overview of Clustering Methods. IOSR Journal of Engineering Apr. 2012, Vol. 2(4), pp. 719-725.
9. Nourahmadi, M. & Sadeqi, H., 2022. Portfolio Diversification based on Clustering Analysis for Tehran Stock Exchange Companies. Iranian Journal of Accounting, Auditing and Finance.
10. Puerto, J., Rodríguez-Madrena, M. & Scozzari, A., 2020. Clustering and portfolio selection problems: A unified framework. Computers & Operations Research, Volume 117.
11. Rani, S. & Sikka, G., 2012. Recent Techniques of Clustering of Time Series Data: A Survey. International Journal of Computer Applications (0975 – 8887) Volume 52– No.15.
12. Vijayalakshmi Pai, G. & Michel, T., 2009. Evolutionary Optimization of Constrained K-Means Clustered Assets for Diversification in Small Portfolios. IEEE Transactions on Evolutionary Computation, 13(5), p. 1030–1053.
13. Wu, D., Wang, X. & Wu, S., 2022. Construction of stock portfolios based on k-means clustering of continuous trend features. Knowledge-Based Systems 252, 109358.
14. Wu, Y.-P., Wu, K.-P. & Lee, H.-M., 2012. Stock Trend Prediction by Sequential Chart Pattern via K-Means and AprioriAll Algorithm. Tainan, Taiwan, s.n., p. 176–181.
15. Yang, H., Lee, H. J., Cho, S. & Cho, E., 2016. Automatic classification of securities using hierarchical clustering of the 10-Ks. Washington DC, USA, s.n., p. 3936–3943.
16. Zhu, Z. & Liu, N., 2021. Early Warning of Financial Risk Based on K-Means Clustering Algorithm. Hindawi, Complexity, Volume 2021.

Legal aspects